NewsSearch: An Architecture for Information Retrieval of Online News

Nuno Maria
Mário J. Silva

DI-FCUL TR-99-3

April 1999

Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
Campo Grande, 1700 Lisboa
Portugal

Technical reports are available at `http://www.di.fc.ul.pt/biblioteca/tech-reports`. The files are stored in PDF, with the report number as filename. Alternatively, reports are available by post from the above address.

NewsSearch: An Architecture for Information Retrieval of Online News

Nuno Maria, Mário J. Silva
FCUL - Faculdade de Ciências da Universidade de Lisboa,
Campo Grande, 1700 Lisboa, Portugal
{nmsm, mjs} @di.fc.ul.pt

Abstract - Efficient information retrieval of highly dynamic information, such as news, is a complex task. As a result, search and retrieval environments for continuously updated news, from other sources than the largest media conglomerates, are almost absent on the Internet. To address this problem, we propose an architecture for a retrieval mechanism designed to improve retrieval efficiency of online news for the networked communities not indexed by the global search engines.

I. Introduction

The explosion in the availability of online information easily accessible trough the Internet is a reality. As the available information increases, the inability to process, assimilate and use such large amounts of information becomes more and more apparent. Online news information suffers from these problems. Currently available tools and search-engines, although sophisticated, present inefficient behavior, as they do not index journalistic and online information sites with the required frequency and effectiveness. These limitations become especially visible for medium-sized national communities with specific information interests, such as the Portuguese web.

If we attempt to use one of the popular search engines to locate Portuguese news, we usually cannot find references to articles less than a month old. In addition, most links to news pages are broken. This happens because online news are very volatile. New information published each day needs to be indexed. In Portugal, we found that only two out of nine online publications provide a search engine on their own archives and only one offers free search and retrieval access to all archived news. All other publications restrict access to their archive, making it impossible to keep accurate links to previously published information. (see Table 1)

Along with this large mass of information, we also have different users with different needs that want to search and retrieve from these archives. Journalists would like to search and review all news, recent and old, from competitor publications when researching a story. Ordinary users often want to query for all recent news from a specific field. These queries bring another factor of complexity to a retrieval environment for dynamic information. How to classify information so those relevant articles can be easily discovered from the huge amount of news published every day?

We believe that the use of intelligent software can help us index specialized information published on the Internet, taking into account its semantics and update constraints. A specialized index for Portuguese news could make use of the knowledge about the update schedule and site organization of the most relevant online publications and make the necessary visits when new information is added. The results of searches on this index could then be combined with those obtained from other search engines to provide more relevant information. In this paper, we present the architectural design of **NewsSearch**, a prototype search environment using this approach as an attempt to overcome current limitations of existing general-purpose search engines.

Portuguese Online Publication	Subject	Available Archive	Search Engine	Periodicity
Correio da Manhã (http://www.correiomanha.pt)	General	Last 3 editions	**No**	Daily
Expresso (http://www.expresso.pt)	General	All Editions	Yes (All editions)	Weekly
Diário de Notícias (http://www.dn.pt)	General	Last 3 editions	**No**	Daily
InforDesporto (http://www.infordesporto.pt)	Sports	N/A	**No**	Daily
Jornal de Notícias (http://www.jn.pt)	General	Last 7 Editions	**N/A**	Daily
O Jogo (http://www.ojogo.pt)	Sports	Last 7 Editions	**No**	Daily
Público (http://publico.pt)	General	Last 7 editions	Yes (in last 7 editions)	Daily
TSF (http://www.tsf.pt)	General (Radio)	All articles	**No**	Daily
TVI (http://www.tvi-online.com)	General (TV)	Last 5 editions	**No**	Daily

(April - 1999)

Table 1. Portuguese major online publications according to subject, available archive, search engine and periodicity.

This paper is organized as follows. In the next section, we identify the deficiencies of general-purpose search engines and related work on the subject. We then present the architectural design of NewsSearch. In section IV, we present an overview of the implementation and discuss how we intend to measure the efficiency of NewsSearch. Finally, we present our conclusions and directions for future work.

II. Indexing on the Internet

In April, 1999, we queried Altavista [16] for all the documents from any of the Portuguese online publications listed in Table 1 dated between February 1, 1999 and April 1, 1999. The answer ranged from "*AltaVista found about 1 Web pages for you*" to "*Altavista found no document matching your query*". Meanwhile, in the same period, about six thousand documents were published in average for each of these publications (assuming one hundred documents per edition per day). Any user looking in this search engine for documents on a particular story would quickly realize that documents published on that story in Portuguese daily online publications had disappeared or were never published. This experience illustrates the current indexing scenery in the online news environment, where update constraints severely limit coverage and use for this general-purpose search mechanism.

Infoseek [17], gives a solution to this problem by extending its coverage and indexing major news web sites across the Internet including REUTERS, Fox News and Washington Post. The online publications scanned on a daily basis provide lots of documents, which are also daily indexed, requiring continuos computing effort.

However, a query to Infoseek for a keyword like "Michael Jordan" will only get recent documents. What about the endless news articles published worldwide about him in the past? For some news locations also indexed by Infoseek, broken links to news two weeks old are common.

All general indexers face the problems described above. Many documents are published in a daily basis not only on Portuguese online publications but worldwide. However, the size of this community is not big enough to get "attention" from large general indexers. In addition, many of the content producers do not archive these documents or do not give the possibility of searching on their corpuses. We are

continuously loosing information, as there is no link consistency or even archiving guaranty of publications by their producers [11].

One approach to this problem could be the creation of a big distributed digital library of news information. This digital library would merge each Internet news web site in a single and uniform system, built as collections of independently developed components that rely on each other to accomplish a larger task. This is the only way for a digital library of news information to scale up to a national or international level. The services for these repositories often need to be operated by independent organizations and the interoperability among these distributed repositories would be the key problem to solve [12].

The NCSTRL (the Networked Computer Science Technical Research Library), a distributed digital library of computer science research reports is a successful example for such a federated library [2]. NCSTRL bases its interoperability in Dienst, a protocol and architecture for distributed digital libraries which has been broadly adopted by a number of research institutions and other distributed collections. The designers of NCSTRL enumerate the principles that should preside in the construction of such a distributed library:

- Open Architecture – The functionality of a digital library system should be available in the form of distinct functional units built with well-know software engineering principles with operational semantics exposed through an open protocol;
- Federation – Such a digital library should be compose of functional units (or services). New services can be added interacting with existing ones using established protocols.
- Distribution – The collections, components of this digital library should be spread over the Internet, but presented to the user as a single uniform system.

Another example of a system based on same architectural guides is the Stanford InfoBus [13]. The InfoBus is an architecture that gives clients uniform access to distributed, heterogeneous information resources and services. InfoBus defines a set of protocols for managing items and collections, metadata, search, payment, and rights and obligations.

The principles above presented, ruled the design of ARIADNE, a digital library for Portuguese news information [10]. ARIADNE is jointly developed by University of Lisbon with “Público”[19], a national daily newspaper. It is a digital publishing infrastructure where all the information used and produced by journalists is organized in a common database. From the information in this library, ARIADNE generates multiple publications in digital format. The system is composed by a set of modules, such as personalized services using information filtering or a payment server for billing the use of information.

ARIADNE positions itself as a member of a federated digital library of news information. However, ARIADNE does not rely, yet, on other similar repositories to provide a full coverage of national online news publications. As a result, ARIADNE indexes the other publications and stores that information in its own repository. For some sites, ARIADNE even collects the entire publication, as there is no guaranty of archival by their publisher. However, with this approach some legal copyright problems are raised due to the widespread uncertainty about legal requirements for managing intellectual property in digital environments [15].

One important service provided by ARIADNE is text categorization or classification, the assignment of natural language texts to one or more predefined category based on their contents. Although many automatic classification algorithms are tested and tuned using news information corpuses (e.g. the Reuters-21578 collection [20]), the automatic classification of news articles, into a set of standard categories, is almost absent in any search engine where news information is available. There is no common agreement for news information classes. Each publication has its particular structure and classification scheme, each of which defines the local category for an article. Even so many automatic classification algorithms for text information have been developed and extensive research results are available [4].

Another problem faced by ARIADNE is the nonexistence of a metadata representation standard for online news information. The Dublin Core Metadata Initiative [3], defines a general set of metadata atributes, called element set, for digital documents. However, news articles have a particular structure not completely covered by this generic element set. In addition, this basic metadata element set is not currently supported by Portuguese online publications in general.

NewsSearch is the functional unit of ARIADNE responsible for information classification and retrieval in the archived corpus. In the next section we present the design architecture for Power Search, and our approach to each of the presented problems.

III. News Search Architecture

NewsSearch is an advanced tool for information retrieval not only inside ARIADNE's digital library (see Figure 1) but also from external news sources. NewsSearch architecture, with all its main components, is presented in Figure 2. Its main modules include the *Index*, the *Classifier* and *Thesaurus*, the *Index Builder* and the *Multi-Search Engine* modules. These components share a common interface bus, used by all of ARIADNE's digital library services. This enables sharing of other resources available in the digital library, such as the multimedia documents repository, and offers services to other digital library components.

A detailed description of the behavior of each NewsSearch component follows.

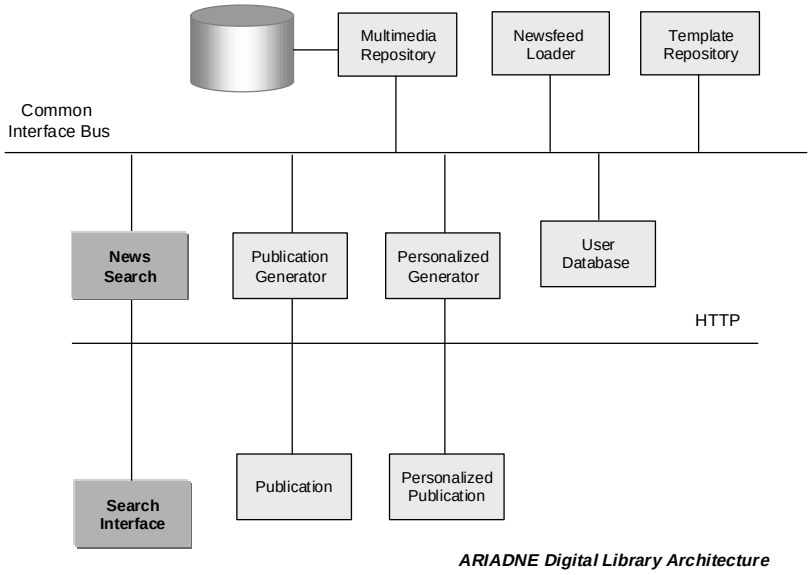

Figure 1. ARIADNE's digital library architecture. Its services include loading news articles and generating electronic publications. NewsSearch is a component of the digital library with a specific interface accessible through a common information bus.

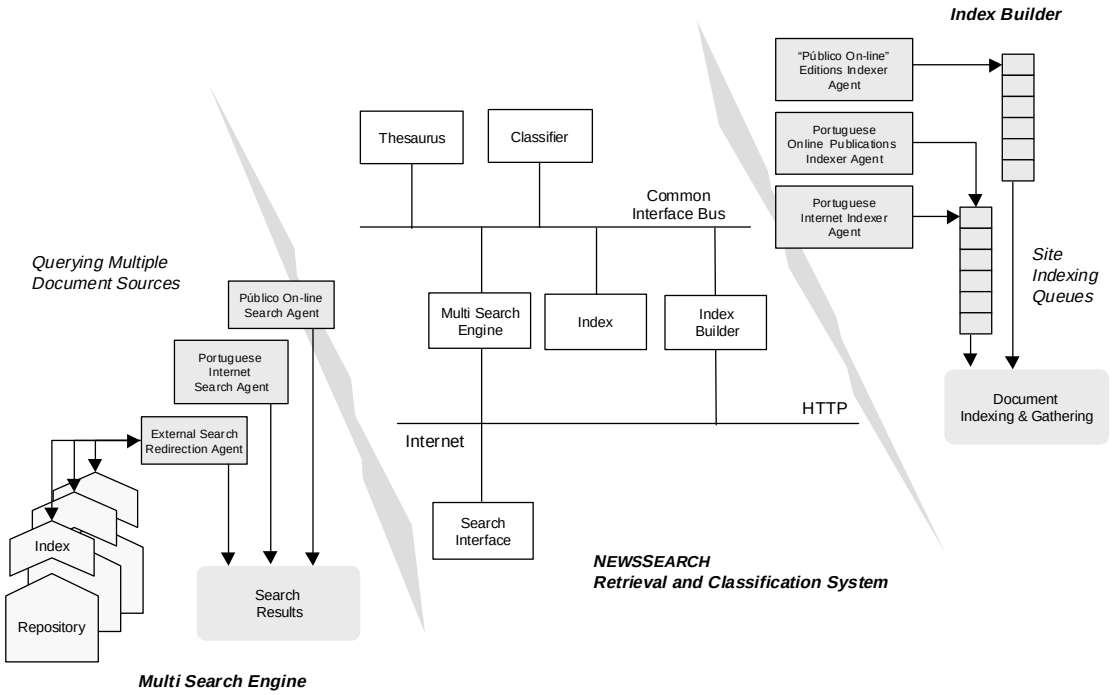

Figure 2. NewsSearch architecture with its main components. These components are interconnected with ARIADNE's common interface bus, enabling the use of other resources in the digital library. The Index Builder defines seperated queues for documents to index based on publications update constraints. The Multi Search Engine involves three kinds of agents in the query process.

The **Classifier** plays an important role, as it categorizes each document, a necessary process to better structure and provide easy discovery of relevant documents in the mass of information gathered by NewsSearch. ARIADNE defines a set of top-level categories in the news environment to overcome the diversity in categorization schemes of different publications, unifying this heterogeneity in fourteen classification topics (see Table 2):

National Politics	Media
International Politics	Education
Health	Science and Technology
Environment	Culture
Work	Sports
Religion	Business and Economics
Local Information	Computers and Internet

Table 2. Top-level categories for text classification in ARIADNE.

This set of top-level categories is common to most European news media. We classify all indexed documents independently of their source or the pre-classification by its publisher according to this scheme.

Given the mass of information to index and its dynamic nature, we had to have automatic classification. NewsSearch's Classifier uses Support Vector Machines – SVM, an algorithm for learning text classifiers from examples [7]. SVM's are very promising as have shown to be very accurate, quick to train, and quick to evaluate [4]. With automatic classification we can easily adapt to new topics and classify information closer to users needs and expectations.

The **Thesaurus** component is used by NewsSearch to expand queries as common in information retrieval engines.

The **Index** is a database of all information gathered by the system. This component holds the keywords, classification obtained by the Classifier module and the location of each document indexed. Location is defined as a link to an outside reference or to the multimedia documents repository of ARIADNE.

The **Index Builder** generates the Index. The top right side of Figure 2 details the architecture of this component. It has two indexing queues managed by a set of software agents [5]. Each queue holds the collection of documents to index by an associated web indexer. One queue is dedicated to documents in the ARIADNE internal repository. The other holds the locations of external publications that need to be indexed. The indexing priority in this queue is determined by the publication periodicity. This information is managed by an agent responsible for the queue organization. After the document is retrieved, the Index Builder requests the classification of the document and saves the references in the Index.

The Index Builder, given its agent-based architecture, can control indexing policies of many collections in a very flexible way. Since this component also retrieves entire publications, the problems of link inconsistencies are minimized, as ARIADNE provides access to its users, when legally possible, to external publications.

The Multi Search Engine seeks pointers for user queries and uses the Thesaurus to expand each query, thus improving precision. Queries are made from the Search Interface and processed in the multi-search engine. This component has three software agents involved in the query process. These agents return to the user the results for the query according to a particular template. A more detailed behavior of the Multi Search Engine is presented on the bottom left side of Figure 2. The agents involved in the query processing include an agent for searching the ARIADNE digital library; an agent for searches in the index created with the information gathered from the Portuguese online publications, and an agent responsible for redirection of searches to external document collections. This last agent is by far the most complex component as it is designed to work with different search engines. In general, these systems only support particular vendor protocols on query evaluation and merging query results. A common meta-searching protocol has been proposed [6], but is only supported by a limited number of search engines. As a result NewsSearch needs to integrate the different protocols of external search engines in a common query protocol.

The **Search Interface** is modeled following the four-phase user interface framework for text searches proposed by Scheidernman *et al.*[14], which emphasizes user-interface clarity and consistency.

The integration of all these components into NewsSearch is our proposed solution to overcome some of the problems introduced in the previous section and detected in present retrieval systems. Given its own specific Thesaurus of news related terms, a user querying NewsSearch for pointers on a set of keywords has a good chance for improving recall – the number of relevant documents retrieved [9]. In addition, this effect is further amplified by the use of the Multi Search Engine. This gives the user freedom to query other search engines transparently, through the use of the external search redirection agent. With this approach, users has a single and uniform interface to a set of repositories and document collections without needing to access different servers and using different interfaces, increasing retrieval effectiveness and user satisfaction.

IV. Implementation

The development of NewsSearch follows an iterative approach with three main stages within each iteration: architectural design, implementation and evaluation. The architectural design of NewsSearch is completed for the first iteration. We are now in the implementation phase evaluating existing components to be integrate and developing other new components according to the design decisions presented in the previous section. After this stage, functionality and performance of NewsSearch will be evaluated. Then, according to the results, design decisions will be refined. This cycle will be iterated until optimal performance is achieved.

For the task of web indexing, several tools are available. We evaluated two: Altavista Search Intranet [17] and Harvest Software [1]. Altavista Search Intranet is a commercial tool designed to index documents on a single Intranet or on sites all over the Internet. This advanced tool as a good performance both on indexing and handling queries. It includes an advanced tool to compose the search interface and a software development kit to manipulate its indexes and structures. However, it has a prohibitive cost for a project of our nature. Harvest software, on the other hand has unrestricted use licensing terms. This tool is not as sophisticated as Altavista but still provides many required functionalities for gathering, extraction, organization, searching, caching, and replication of relevant information across the Internet. Another key advantage of Harvest relies on the availability of configuration and easy integration with other components since its source code is freely available. This lead us to decide for the Harvest software as the component of NewsSearch we build on for web indexing.

The evaluation process is the most complex stage in this development. The structure and algorithms chosen largely dictate the performance of an information retrieval system. However, the evaluation of that performance belongs to the user. Each user will judge system performance from a distinct viewpoint, and many characteristics contribute to the user's perception of system effectiveness. These include speed, the accuracy of the response to a query, and the amount of irrelevant information included in any response.

So, the only way to evaluate NewsSearch is to submit it to a test with a number of users and analyze their reactions and satisfaction as the prototypes are made available. Koenemann and Belkin, have conducted an empirical study similar to that we will make [8]. ARIADNE currently have ten thousand daily users, and NewsSearch as a part of ARIADNE will probably have the same number of users. This large audience will then be used as a test platform to evaluate NewsSearch.

V. Conclusion and Future Work

Current general search and retrieval mechanisms available on the Internet present inefficient behavior to users in small communities with specific information interests. Users interested in online news of small nations are an example of such communities.

With NewsSearch, we try to overcome these limitations of frequency and coverage, providing one common and uniform interface for searching in different collections and distributed repositories, hoping to raise search effectiveness and user satisfaction for that information that is not properly handled by the large search engines of the Internet.

As Future work, in addition to completing the development and evaluation of NewsSearch, we plan to integrate a relevance feedback mechanism to improve retrieval effectiveness along with user satisfaction.

VI. References

[1] Bowman C., Danzig P., Hardy D., Manber U., and Schwartz M.: The Harvest information discovery and access system, *in Proceedings of the Second International WWW Conference '94: Mosaic and the Web*, 1994;

[2] Davis, J. and Lagoze C.: NCSTRL: Design and Deployment of a Globally Distributed Digital Library, *in IEEE Computer*, February 1999;

[3] Dublin Core Metadata Initiative, http://purl.oclc.org/metadata/dublin_core;

[4] Dumais S., Platt, J., Heckerman, D. and Sahami M.: Inductive Learning Algorithms and Representations for Text Categorization, *in CIKM '98, Conference Proceedings on Information and Knowledge Management*, 1998;

[5] Genesereth M. and Ketchpel S.: Software Agents, *in Communications of the ACM Vol.37 Nº7*, July 1994;

[6] Gravano L., Chang C., Paepcke, A. and García-Molina H.: STARTS: Stanford Proposal for Internet Meta-Searching, *in SIGMOD '97, Conference Proceedings on Management of Data*, 1997;

[7] Joachims T.: Text Categorization with Support Vector Machines: Learning With Many Relevant Features, *in Proceedings of the European Conference on Machine Learning*, 1997;

[8] Koenemann J. and Belkin N.: A Case for Interaction: a Study of Interactive Information Retrieval Behavior and Effectiveness, *in CHI '96, Conference Proceedings on Human Factor in Computing Systems*, 1996;

[9] Korfhage R.: Information Storage and Retrieval. Jonh Wiley & Sons, 1997;

[10] Maria, N., Gaspar, P., Grilo, N., Ferreira A., Silva, M.: ARIADNE - Digital Library Architecture, *in Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, 1998;

[11] Maria, N., Gaspar, P., Ferreira A., Silva, M.: Information Preservation in ARIADNE, *in Proceedings of the 6th DELOS Workshop*, 1998;

[12] Paepcke, A., Chang C., García-Molina H. and Winograd T.: Interoperability for Digital Libraries, *in Communications of the ACM Vol.41 Nº4*, April 1998;

[13] Roscheisen M., Baldonado M., Chang C., Gravano L., Ketchpel S. and Paepcke, A.: The Stanford InfoBus and Its Service Layers: Augmenting the Internet with Higher-Level Information Management Protocols, *in Digital Libraries in Computer Science: The MeDoc Approach, Lecture Notes in Computer Science No. 1392*, Springer, 1998;

[14] Shneiderman, B., Byrd D. and Croft B.: Sorting Out Searching: a User-Interface Framework for Text Searches, *in Communications of the ACM Vol.41 Nº4*, April 1998;

[15] Samuelson P.: Good News and Bad News on the Intellectual Property Front, *in Communications of the ACM Vol.42 Nº3*, March 1999;

[16] Altavista, http://www.altavista.com;

[17] Altavista Search Software, http://altavista.software.digital.com;

[18] InfoSeek, http://infoseek.go.com;

[19] Público On-line, http://publico.pt;

[20] Reuters-21578 collection, http://www.research.att.com/~lewis/reuters21578.html;